INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Application de modèles de langages à la reconnaissance de promoteurs

Jean-Yves GIORDANO

N° 3099

janvier 1997

THÈME 3

ISSN 0249-6399

INRIA
RENNES

Application de modèles de langages à la reconnaissance de promoteurs

Jean-Yves GIORDANO

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Repco

Rapport de recherche n° 3099 — janvier 1997 — 37 pages

Résumé : Nous nous intéressons dans ce rapport aux possibilités d'application d'outils de modélisation du langage naturel à l'analyse de séquences génomiques. Le problème biologique est la modélisation de régions promotrices eucaryotes, dans un but de prédiction et d'explication. Une partie de ce rapport traite de la représentation des *sites de fixation*, qui jouent un rôle déterminant dans la régulation de l'expression des gènes. L'autre partie concerne la caractérisation des régions promotrices sans connaissances *a priori*.

Plusieurs approches ont été tentées, parmi lesquelles l'inférence grammaticale, les modèles de Markov cachés, les ngrams et multigrams. Les résultats obtenus sur les sites de fixation sont comparables à ceux obtenus dans la littérature sur le sujet, et nous semblent difficilement exploitables dans un but de caractérisation, en raison de la complexité des mécanismes biologiques liant ces sites à la régulation. Les résultats concernant la caractérisation des régions promotrices sont probants mais révèlent une différence fondamentale entre langages naturel et biologique. Alors que les textes en langage naturel sont caractérisés par une grande stabilité des fréquences des différentes combinaisons de lettres, des régions fonctionnellement similaires de l'ADN peuvent être d'une grande diversité quant à leur composition. Nous pensons que la prise en compte de cette variabilité est essentielle pour l'application de modèles de langages à l'ADN. Nous décrivons une première tentative dans ce sens, en adaptant une méthode simple de comparaison des fréquences. Les résultats sont concluants et suggèrent que l'adjonction de telles corrections aux techniques plus sophistiquées existantes puisse être d'un grand intérêt.

Mots-clé : analyse de séquences, génome, modèles de langages, promoteurs, inférence grammaticale, modèles de Markov cachés, ngrams, multigrams

(Abstract: pto)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 02 99 84 71 00 – Télécopie : (33) 02 99 84 71 71

Application of language models to promoter recognition

Abstract: This document deals with the application of classical language modeling tools to a DNA recognition problem. Our aim is to model eucaryotic promoter regions, the purpose being to be able to predict whether a sequence is a promoter or not, as well as to infer informations about their mode of action. The first part of the document deals with the representation of DNA binding sites which play a determinant role in the regulation of transcription. The second part deals with modeling promoter regions as a whole, using no biological knowledge.

Several approaches are explored: grammatical inference, Hidden Markov Models, ngrams and multigrams. The results obtained concerning binding sites are comparable with those obtained with the commonly used method. The mechanisms that link binding sites to regulation are complex and yet relatively obscure, so that it seems difficult to characterize promoters from the knowledge of binding sites only. The results concerning the characterization of promoter regions from scratch are satisfying but reveal a first fundamental difference between natural language and genetic texts. Indeed, while the frequencies of the different combinations of symbols are remarkably conserved in natural language ([Dam95]), the composition of functionally equivalent sequences of DNA can vary widely. We think that taking into account this variability is essential when trying to apply natural language models to DNA. The results of a first attempt based on this consideration are conclusive and suggest that adapting the existing techniques to the variability of DNA composition might be of great interest.

Key-words: sequence analysis, genome, models of languages, promoters, grammatical inference, Hidden Markov Models, ngrams, multigrams

Table des matières

1 Introduction

Nous présentons ici les notions biologiques employées et exposons le problème de la reconnaissance des promoteurs. L'analogie entre les séquences génomiques et les langages naturels (il s'agit dans les deux cas de séquences de symboles) incite à transposer les techniques de modélisation du langage à ce problème. Nous rappelons brièvement le principe de quatre techniques étudiées : l'inférence grammaticale et les modèles de Markov cachés pour les sites de fixation, les ngrams et les multigrams pour la modélisation des régions promotrices.

1.1 Problème de la reconnaissance des promoteurs

Nous définissons ici les termes employés concernant l'expression des protéines et leur régulation chez les organismes eucaryotes (dont les cellules présentent un noyau à l'intérieur duquel se trouve la molécule d'ADN). Le cas des procaryotes (essentiellement des bactéries) est plus simple, et la structure des régions promotrices est beaucoup mieux conservée.

Les séquences biologiques sont des macro-molécules constituées par l'enchaînement d'acides appartenant à un éventail réduit de molécules possibles. Il existe essentiellement 3 types de séquences biologiques, que l'on distingue en fonction de ce "vocabulaire" (de 4 à 20 lettres) sur lesquelles elles sont construites : ADN, ARN et acides aminés (protéines).
ARN et proteines sont codés et donc formés à partir d'une matrice d'ADN. Plus spécifiquement, l'expression des protéines à partir de la molécule d'ADN s'effectue chez les eucaryotes en trois étapes. Dans un premier temps, un fragment d'ADN est transcrit (copié) pour former une molécule d'ARN. Dans un deuxième temps, plusieurs parties de la molécule d'ARN sont supprimées (épissage). Les séquences de l'ADN initial correspondant à l'ARN résultant sont appelées des exons. La dernière étape consiste à traduire la molécule d'ARN pour produire une protéine.

Ce mécanisme est général. Cependant, les besoins en protéines variant suivant les tissus, un mécanisme de régulation intervient sur leur expression. Il

existe sur le gène, en amont de chaque ensemble de nucléotides à transcrire, une région dite promotrice chargée d'activer ou d'inhiber la transcription. Sur la région promotrice viennent se fixer des protéines (dites facteurs de transcription) facilitant ou inhibant la transcription[1]. Les endroits où se fixent ces protéines sont les sites de fixation.

Alors que le problème de la détection d'exons a fait l'objet de nombreuses études ayant mené à des algorithmes relativement fiables (voir par exemple [YXMSU94]), les connaissances sur les mécanismes de régulation sont limitées. Il n'existe que peu d'algorithmes de reconnaissance des promoteurs et ils présentent l'inconvénient majeur de produire un nombre élevé de *faux positifs*, i.e. de séquences non promotrices reconnues promotrices. Le taux de faux positifs est un facteur important, étant donné le nombre restreint de régions promotrices (que l'on peut estimer à 100000 pour le génome humain, la taille d'une région étant de 1000 nucléotides), comparé à la taille du génome (3.10^9 nucléotides pour le génome humain).

1.2 Approche proposée

Nous nous sommes proposé d'examiner les possibilités d'application à ce problème de techniques classiques en traitement du langage naturel (aux niveaux phonétique, lexical et syntaxique). Présentons quatre techniques développées par la suite : l'inférence grammaticale et les modèles de Markov cachés pour les sites de fixation, les ngrams et les multigrams pour la région promotrice.

- **Inférence grammaticale**

  L'inférence grammaticale consiste à inférer la syntaxe sous-jacente à un langage, sous forme d'un langage formel (la classe de langages généralement inférée est la classe des langages réguliers). Elle peut être utilisée

1. la région que nous appelons promotrice est plus exactement constituée d'une zone promotrice au sens strict (longue d'une soixantaine de nucléotides et enserrant le début de transcription), dont le rôle est essentiellement de rendre possible la transcription, et d'une zone située en amont et dite régulatrice, qui elle peut à la fois promouvoir ou inhiber la transcription. Cette distinction étant secondaire pour la suite de l'exposé, nous confondrons promotion et régulation.

pour l'acquisition automatique de modèles de langages permettant de guider la reconnaissance de la parole, au niveau phonétique [GSCT94] ou syntaxique [Dup96], aussi bien qu'en reconnaissance de formes quand une partie des données peut être représentée sous forme de séquences de symboles [Luz94].

– **Modèles de Markov cachés**

Un modèle de Markov caché est un modèle statistique constitué d'un ensemble d'états et de transitions. à chaque état (ou transition) est liée une probabilité d'émission (d'un symbole, d'un vecteur acoustique...). à chaque transition est liée une probabilité de passage par cette transition. Les probabilités associées aux états et transitions ainsi que l'état courant sont inconnues de l'observateur qui n'a connaissance que des séquences émises. à partir de celles-ci les probabilités peuvent être estimées selon differents critères, le but étant l'obtention d'un modèle pour la production de ce type de séquences.

Les modèles de Markov cachés sont utilisés comme modèles acoustiques dans la majorité des systèmes de reconnaissance de la parole, et sont utilisés en bioinformatique, notamment pour la modélisation de la structure des protéines [AKMSH94, Sea96].

– **Ngrams**

Les ngrams sont un modèle de langage statistique intervenant en reconnaissance de la parole au niveau lexical, et permettant de réduire le taux d'erreur du système par le biais d'une estimation de la probabilité *a priori* d'une séquence de mots [Jel90]. Concrètement, la probabilité des séquences de mots est estimée proportionnellement à leur fréquence d'apparition dans un corpus d'apprentissage, la probabilité d'un mot dans une séquence dépendant de ses $n-1$ prédécesseurs, où n est un paramètre fixé.

– **Multigrams**

Ce modèle récent, contrairement aux ngrams, porte sur des suites de symboles de longueur variable. Le principe en est la segmentation d'une sé-

quence de symboles en sous-unités (phonétiques, lexicales ou syntaxiques), la probabilité d'une segmentation étant estimée après entraînement sur un corpus. Un des attraits de cette technique est la production d'unités élémentaires significatives (au moins dans le cas du langage naturel [DB95]).

Ce document décrit les trois étapes de notre travail.

Nous nous sommes attaché dans un premier temps à la description des sites de fixation (à l'aide de grammaires ou de modèles de Markov).

La caractérisation des régions promotrices dans leur ensemble est ensuite abordée à travers le modèle ngrams. Les résultats obtenus dépendant fortement de la composition des séquences, ce modèle est comparé à une méthode simple basée sur la comparaison des fréquences de n-uplets [Dam95]. Les résultats sur le modèle multigrams, bien qu'incomplets (le mode d'apprentissage le plus simple est étudié), confirment que l'application de modèles de langages au problème de la reconnaissance de promoteurs est limitée par le biais introduit par la composition des séquences.

Une première tentative concluante pour se dégager de ce biais est alors décrite (on compare les différences entre les fréquences observées et les fréquences attendues des n-uplets).

2 Description des sites de fixation

Les sites de fixation sont de courtes séquences situées sur la région promotrice, sur lesquelles viennent se fixer des protéines (facteurs de transcription) facilitant ou inhibant la transcription.

Donnons à titre d'exemple quelques sites correspondant au facteur YY1 [SYSGG95] :

```
CAGAGACACAGACGCCAT
         TACAGCCATTATTCCCCA
CAGACTACAATCTACCAT
     TGACCGGCGCCATTGTTA
         GTACGCCATTATCCCTTG
    CAGAAAGGCGCCATTTCG
TTAAAGCTATCCCACCAT
          TTCGCCATTTGCGTGAGT
   TACTGTAATCGCCATACT
     CACACGCGGCCATTTTCC
     AGGGGTCGTCCATTTTTA
TAACACCACCTCCGCCAT
TAACAAACACCCGGCCAT
TGAATCTGCGGCAACCAT
GGGAAACACTCCCGCCAT
          TACACCATATTGCCTCAC
       CAGTACGCCATTACCCTA
         TACGTCCATATTGATTTC
    CATATAGCCGCCATTTAC
```

On remarque que ces séquences ont toutes en commun la sous-séquence CCAT caractéristique de ce facteur. Une telle sous-séquence (appelée *core*) existe pour tous les facteurs, mais elle est rarement aussi bien conservée et son contexte (les nucléotides voisins) joue un rôle important dans l'affinité du site pour le facteur correspondant.

2.1 Méthode des matrices

La méthode classique pour mesurer l'affinité d'un site pour un facteur est la méthode des matrices. Une fois les sites connus alignés autour du *core*, la matrice est construite en fonction du nombre de nucléotides présents à une position donnée. Une matrice correspondant au facteur YY1, calculée sur 53

séquences, est ainsi :

	-5	-4	-3	-2	-1	CCAT	1	2	3	4	5
G	28	45	23	21	79		21	5	2	26	19
A	36	25	17	24	4		19	21	2	15	21
T	26	21	15	4	15		45	70	83	42	30
C	10	9	45	51	2		15	4	13	17	30

à chaque nouvelle séquence contenant le *core* est alors associé un score, somme des scores de chacun des nucléotides la composant. Il représente l'affinité de la séquence pour le facteur, un haut score indiquant un site de fixation potentiel.

2.2 Recherche d'une syntaxe des sites de fixation

Nous avons tenté de découvrir si la séquence de nucléotides constituant un site de fixation était organisée selon des règles syntaxiques simples. L'outil le plus adapté à cette tâche est l'inférence grammaticale, dont l'objet est la production d'une grammaire à partir d'un ensemble d'exemples (éventuellement de contre-exemples). Deux algorithmes ont été testés :

- RPNI (Regular Positive and Negative Inference) [OG92]

  Cet algorithme produit en temps polynomial un automate déterministe minimal compatible avec deux ensembles disjoints de mots I^+ et I^-, c'est-à-dire acceptant tous les mots de I^+ et rejetant ceux de I^-.

- TGI (Tabu Grammatical Inference) [Gio96]

  TGI cherche à produire une grammaire régulière discriminant au mieux deux ensembles de mots, en utilisant une technique classique d'optimisation combinatoire (Tabu Search).

Les contre-exemples utilisés pour ces deux algorithmes sont des séquences extraites de la banque EMBL (European Molecular Biology Laboratory), de longueur égale à celle des exemples et ne contenant pas la sous-séquence CCAT.

Nous avons cherché dans un premier temps à inférer une grammaire à partir de la séquence entière des sites de fixation. RPNI produit l'automate reconnaissant tous les mots contenant CCAT, et seulement ceux-ci. Le résultat est prévisible pour cette méthode qui cherche à inférer un automate minimal. TGI produit des automates complexes, dont les taux de reconnaissance (pour les meilleurs d'entre eux) sont de l'ordre de 95% pour les exemples et 5% pour les contre-exemples. Malheureusement, le taux de faux positifs sur un ensemble de contre-exemples test est de l'ordre de 70%, situation caractéristique d'un apprentissage "par cœur", c'est à dire se contentant de mémoriser sous une forme plus ou moins directe les instances d'entraînement. Ceci conduit à des automates très spécifiques, incapables de détecter une éventuelle structure commune dans les exemples.

Nous avons tenté d'isoler l'information concernant le contexte du *core* en lançant les deux programmes sur les préfixes et suffixes de CCAT. Les résultats obtenus avec RPNI comme avec TGI ne sont pas probants (automates complexes, trop de faux positifs).

Ces premières expériences semblent indiquer que la classe des langages réguliers, parmi les plus simples dans la hiérarchie de Chomsky, ne permet pas de représenter le langage des séquences correspondant à des sites de fixation (si ce n'est en extension). Nous nous sommes interessés en conséquence à des modèles plus statistiques.

2.3 Représentation sous forme de modèles de Markov

La méthode la plus courante pour mesurer l'affinité d'un site pour un facteur est la méthode des matrices (cf. section 2.1). Nous avons tenté de rendre compte de façon plus précise de cette affinité en utilisant un outil plus élaboré. Notre choix s'est porté sur les modèles de Markov cachés, déjà utilisés pour modéliser la structure primaire (i.e. la suite de nucléotides) d'une famille de séquences [AKMSH94, Sea96].

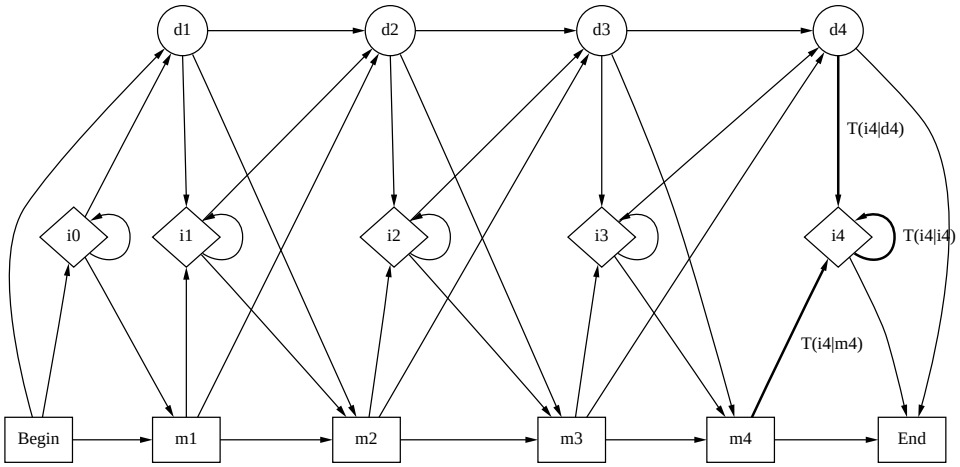

FIG. 1 – *Structure du modèle inféré*

Nous avons utilisé une batterie d'algorithmes du domaine public[2], et remarqué une forte corrélation des résultats obtenus sur un échantillon test avec ceux obtenus par la méthode des matrices.

La structure du modèle inféré, représentée en figure 1, est classique dans le domaine. Le graphe est linéaire et comprend trois types d'états représentant l'insertion (Insert), la suppression (Delete) ou l'émission (Match) d'un symbole. Les états correspondant à une suppression n'émettent pas de symbole, au contraire des deux autres.

Le modèle construit à partir des 53 séquences disponibles compte 12 états d'émission. Comme pour la méthode des matrices, une région entière peut alors être parcourue à la recherche de sites potentiels. Afin de comparer les deux méthodes nous considérons une séquence de 73308 nucléotides, contenant le gène de la β-globine humaine et 350 *core* CCAT. La figure 2 montre la corrélation entre les scores calculés avec la méthode des matrices et le modèle de Markov. Cette corrélation s'explique par le fait que la modélisation matricielle est fina-

2. http://genome.wustl.edu/eddy/hmm.html

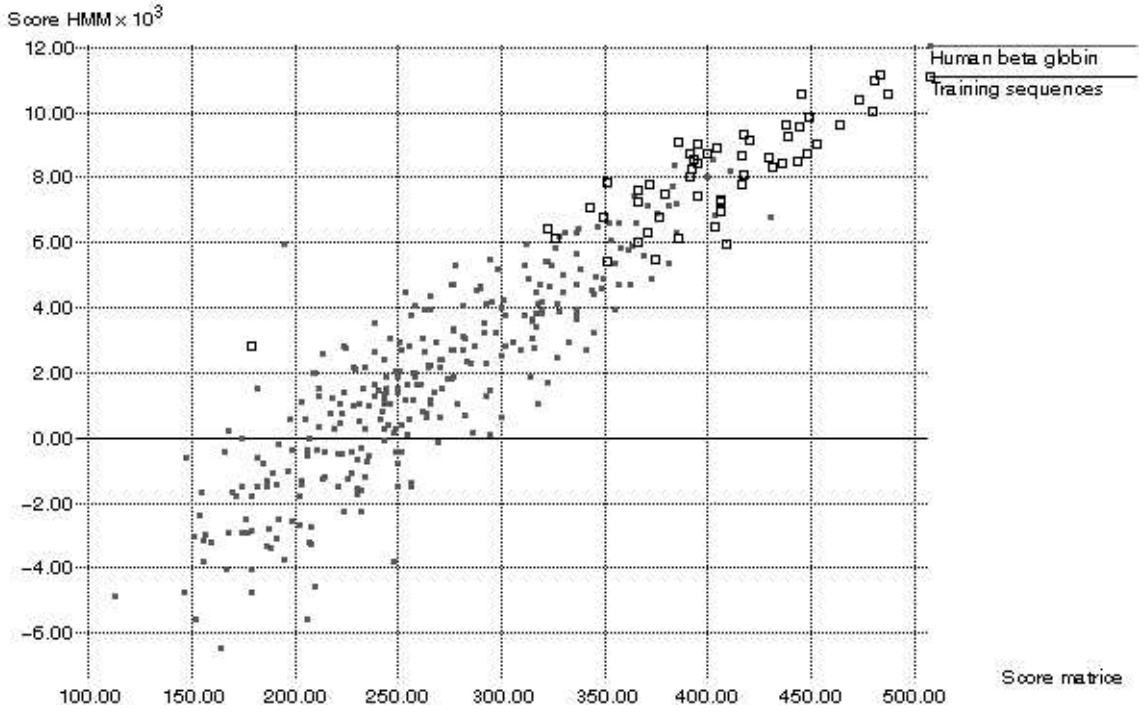

FIG. 2 – *Corrélation Markov/Matrices*

lement équivalente à une chaine de Markov simple.

Il est difficile de déterminer quelle approche est la meilleure (dans le sens où elle permet de mesurer l'affinité d'un site pour un facteur) car la vérification expérimentale est complexe, dépend des conditions de l'expérience (*in vitro, in vivo*) et en dernier ressort de la région de l'ADN entourant le site.

Le mécanisme liant sites de fixation, facteurs de transcription et régulation est par ailleurs complexe et relativement mal connu, et l'affinité intrinsèque d'un site n'est pas le seul critère à prendre en compte. Citons à cet égard un ouvrage de référence [Lew95] : "Any protein that is needed for the initiation of transcription, but which is not itself part of RNA polymerase, is defined as a transcription factor. Many transcription factors act by recognizing cis-acting sites that are classified as comprising parts of promoters or enhancers. We must recognize, however, that binding to DNA is not necessarily the only means of action for a transcription factor. A factor may recognize another factor, or may recognize RNA polymerase, or possibly may be incorporated into an initiation complex only in the presence of several other proteins. The ultimate test for membership of the transcription apparatus is functional: a protein must be needed for transcription to occur at a specific promoter or set of promoters."

En raison d'une part de la difficulté de calculer l'affinité d'un site pour un facteur, d'autre part du mécanisme subtil liant sites potentiels et régulation, nous nous sommes tourné pour la reconnaissance des régions promotrices vers une approche plus globale n'exploitant aucune connaissance sur les classes de séquences considérées.

3 Reconnaissance des régions promotrices

Cette section traite de l'utilisation de modèles de langages statistiques pour la caractérisation de régions promotrices, sans connaissances *a priori*. Le but est dans un premier temps de fournir un outil permettant de parcourir une large séquence à la recherche de promoteurs, ou de déterminer si une séquence donnée est promotrice (prédiction). Nous espérons dans un deuxième temps

être capables de déduire de l'examen des modèles des éléments caractéristiques des promoteurs et susceptibles de nous instruire sur leur mode d'action (explication).

Nous étudierons les modèles ngrams et multigrams. Leur but premier en traitement de la parole n'est pas la discrimination de deux classes de séquences, mais ils constituent des modèles pouvant éventuellement mieux rendre compte des propriétés statistiques des régions promotrices que des méthodes plus simples.

La démarche est la suivante pour les deux approches : deux modèles sont construits pour chaque classe (ADN quelconque ou promoteur). Chacun permet d'associer à une nouvelle séquence une probabilité d'appartenance à la classe qu'il représente. Les critères d'évaluation du modèle seront d'une part le taux de promoteurs reconnus, d'autre part le taux de faux positifs.

Nous constatons dès l'étude du modèle ngrams une forte corrélation entre le score associé à une séquence et la proportion de G et de C dans cette séquence (que l'on désignera par la suite par taux de GC). Nous confrontons alors la méthode ngrams à une méthode simple basée sur les fréquences de n-uplets. Les résultats des deux méthodes sont corrélés et liés aux taux de GC, mais la comparaison est en faveur des ngrams.

Contrairement aux deux méthodes précédentes, le modèle multigrams ne repose pas directement sur un comptage des n-uplets, et présente l'avantage d'opérer une segmentation des séquences. Il est difficile de comparer cette méthode aux deux précédentes, l'ensemble d'apprentissage étant différent, mais on constate pareillement une corrélation entre le taux de GC et le score associé aux séquences.

3.1 Ngrams et comparaison de fréquences

Cette section a pour objet d'évaluer les performances de l'approche ngrams et de la comparer à une méthode simple basée sur la comparaison des fréquences des n-uplets dans les ensembles d'apprentissage d'une part et la sé-

quence à classer d'autre part.

Deux facteurs jouant sur la qualité des résultats sont examinés. Le premier est la taille des n-uplets utilisés : $n = 6$ s'avère être un bon compromis dans notre cas. Un deuxième facteur est la taille des séquences considérées : les prédictions sur des séquences courtes sont de moins bonne qualité que sur des séquences plus longues, mais les performances tendent à se stabiliser à partir d'une certaine taille.

Deux exemples de détection sont donnés en fin de section.

3.1.1 Nature des échantillons

Nous avons utilisé pour les promoteurs 1265 séquences extraites de la banque EMBL, contenant un début de transcription, de longueur variant entre 150 et 1050 nucléotides et situées dans la région [-1000,+50] autour du début de transcription.

381 de ces séquences ont été validées par vérification du début de transcription. Elles totalisent 292068 nucléotides et constituent l'échantillon d'apprentissage positif. Les 884 séquences restantes, non vérifiées, constituent l'échantillon de validation positif.

L'échantillon d'apprentissage négatif est l'ensemble des séquences d'ADN humain d'EMBL (environ 2.10^7 nucléotides). Le poids des promoteurs présents dans cet échantillon étant négligeable, nous avons choisi de ne pas les oter de la base de contre-exemples. L'échantillon de validation négatif est constitué de 24 séquences de longueur supérieure à 50000 nucléotides (au total $1,7.10^6$ nucléotides), qui seront découpées en tronçons de la longueur désirée.

3.1.2 Ngrams

Nous présentons dans cette section la méthode ngrams, et montrons que les résultats dépendent de la composition des séquences.

Modélisation des échantillons

Les ngrams permettent de modéliser une classe de séquences en associant à chaque symbole sa probabilité d'occurrence en fonction de ses $n-1$ prédécesseurs. La vraisemblance d'une séquence $w = w_1 w_2 ... w_N$ $(N > n)$ est alors calculée de la façon suivante :

$$\mathcal{L}(w) = P(w_n|w_1 \dots w_{n-1})P(w_{n+1}|w_2 \dots w_n) \dots P(w_N|w_{N-n+1} \dots w_{N-1})$$

Les probabilités $P(w_i|w_{i-n+1} \dots w_{i-1})$ constituent les paramètres du modèle, et sont estimées d'après les ensembles d'apprentissage.

La vraisemblance associée à une nouvelle séquence par chacun des modèles permet alors de déterminer sa classe d'appartenance.

Résultats

Nous avons calculé deux modèles de 6-grams à partir des échantillons d'apprentissage, et associé à chaque séquence $w = w_1 w_2 \dots w_N$ des échantillons de validation un score $S(w)$:

$$S(w) = \frac{log(\mathcal{L}_{pro}(w)/\mathcal{L}_{dna}(w))}{N - n + 1}$$

w est ainsi reconnu comme promoteur ssi $S(w) > 0$. De plus, le calcul de $S(w)$ prend en compte la longueur de la séquence, de telle manière que des séquences de tailles différentes peuvent être comparées : si $S(w_1) > S(w_2)$ pour deux séquences w_1 et w_2, alors w_1 est un meilleur candidat promoteur que w_2.

La figure 3 montre la corrélation entre $S(w)$ et le taux de GC dans w. Les séquences de l'échantillon négatif sont de longueur 600. 40% des promoteurs sont reconnus comme tels, ainsi que 6,3% des séquences d'ADN (ainsi que nous le verrons, certaines de ces dernières sont effectivement des promoteurs).

La prédiction est directement liée au taux de GC, ce qui s'explique par le fait que les séquences promotrices sont en moyenne plus riches en GC que le reste de l'ADN. Ceci nous amène à confronter cette méthode à une simple comparaison des fréquences des n-uplets.

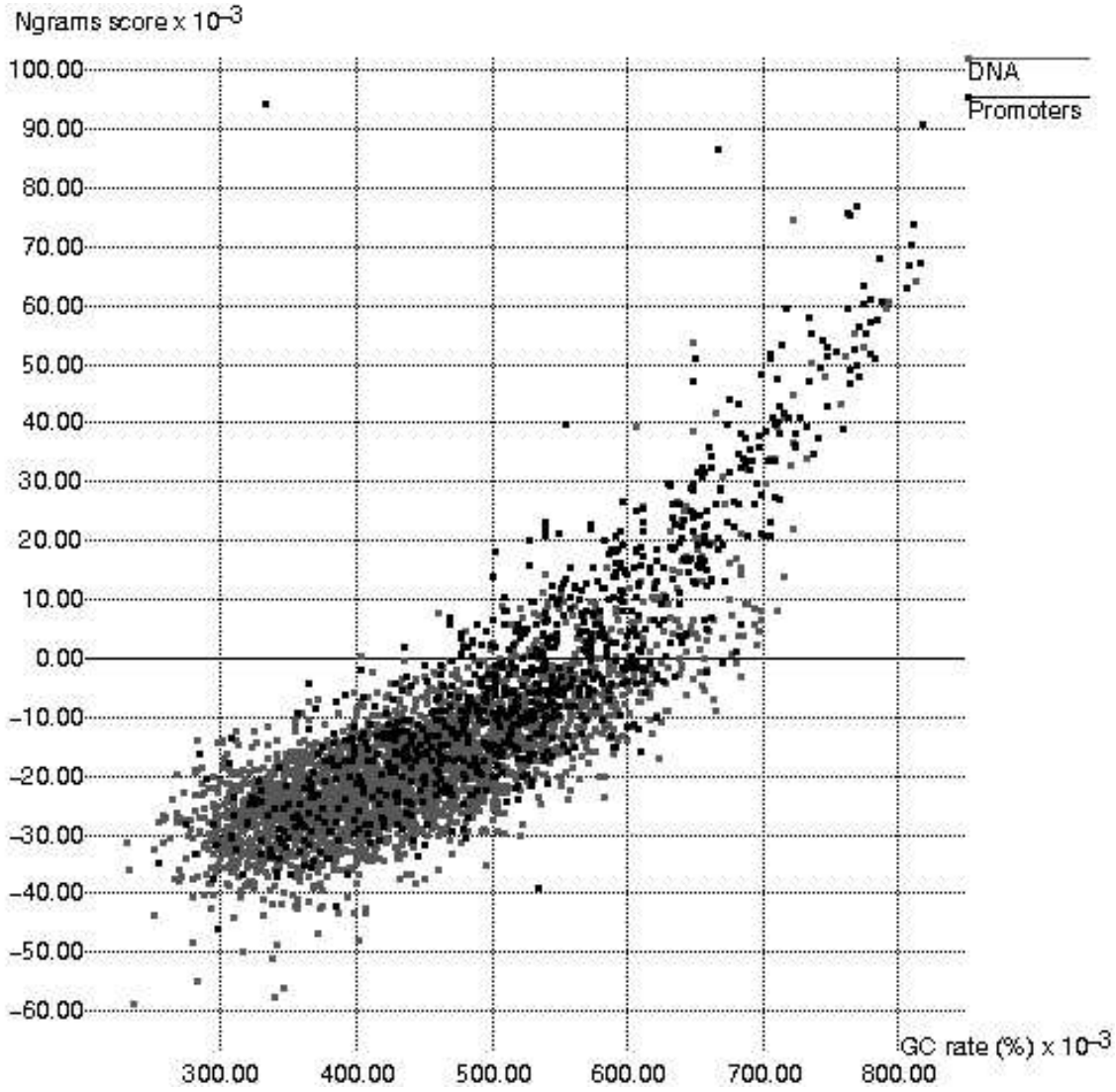

FIG. 3 – *Corrélation $S(w)$/Taux de GC*

3.1.3 Comparaison des fréquences des n-uplets

Les paramètres des modèles ngrams sont calculés par comptage des n-uplets des ensembles d'apprentissage. Nous présentons une autre méthode basée sur le comptage des n-uplets, proposée par M. Damashek [Dam95] pour la reconnaissance et la classification de textes écrits en langage naturel (il s'agit de regrouper différents textes par famille de langue, par langue ou sujet traité), et testée pour la reconnaissance de promoteurs dans [Leg95].

Principe de la méthode

Pour une séquence donnée les fréquences $f_i (i = 1 \ldots 4^n)$ de tous les n-uplets sont calculées. Ces fréquences correspondent à un vecteur dans un espace de dimension 4^n. Puisque la somme des fréquences est égale à 1, ce vecteur est contenu dans l'hyperplan d'équation :

$$\Sigma_{i=1}^{4^n} f_i = 1$$

Une séquence peut donc être représentée par un vecteur. L'avantage de cette représentation est qu'une distance entre deux séquences peut être calculée, comme la valeur de l'angle entre leurs vecteurs associés.

De la même façon, un ensemble de séquences est représenté par un vecteur dont les coordonnées sont les fréquences d'apparition des n-uplets dans cet ensemble. étant donnés deux ensembles de séquences, deux vecteurs caractéristiques peuvent être calculés. La classification d'une nouvelle séquence s'effectue alors par comparaison du vecteur la représentant et de ces deux vecteurs caractéristiques.

Résultats

La figure 4 montre que les deux méthodes tendent à classer les séquences de la même manière (n vaut 6 dans les deux cas, les séquences de l'échantillon négatif sont de taille 600).

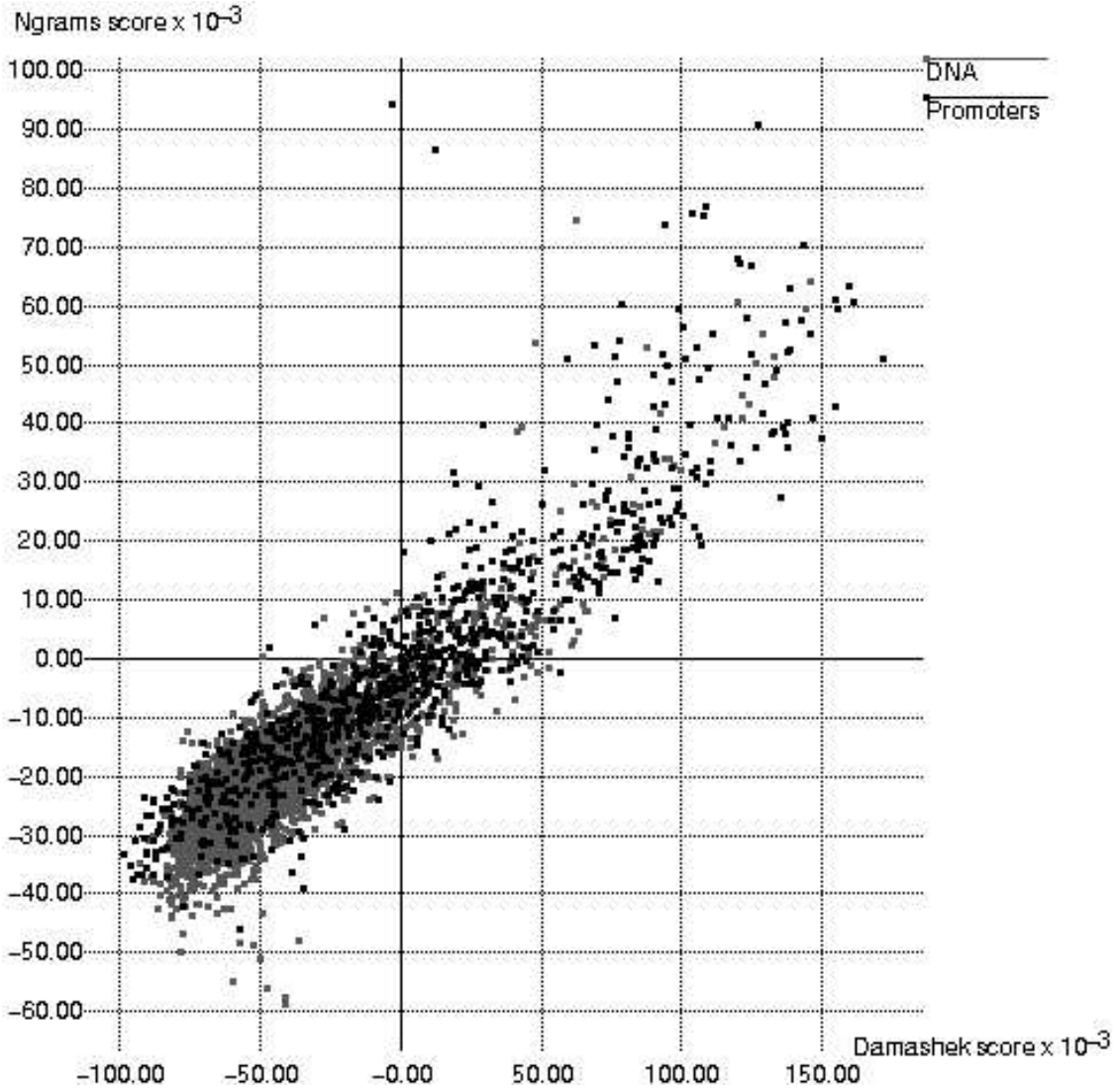

FIG. 4 – *Corrélation Ngrams/Damashek*

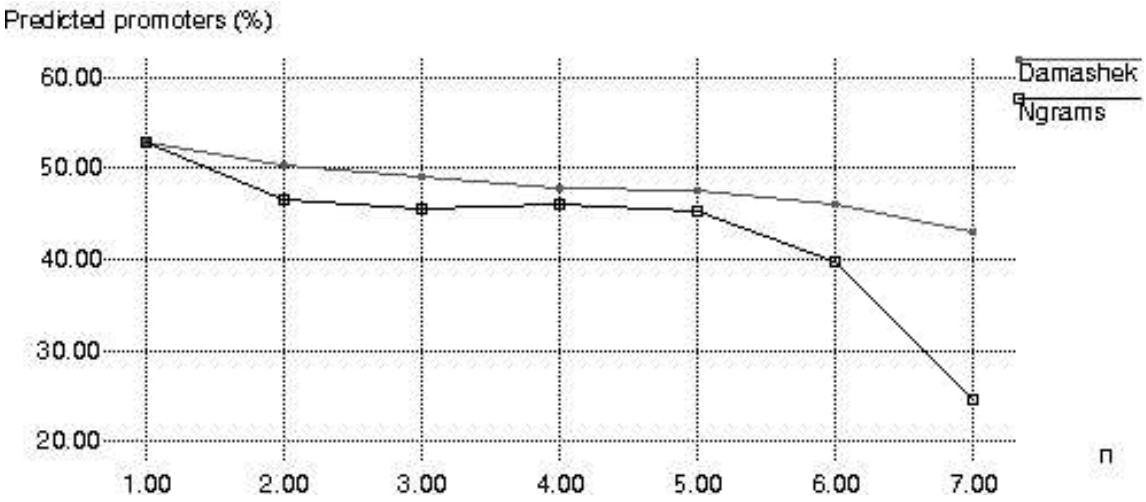

FIG. 5 – *Taux de vrais positifs* ($n = 1 \ldots 7$)

Ce résultat n'est pas surprenant, les paramètres des modèles ngrams étant estimés par une procédure de comptage des n-uplets. Les deux méthodes diffèrent toutefois quant aux taux de reconnaissance, comme le montre la section suivante.

3.1.4 Comparaison des deux méthodes

La comparaison qui suit s'oriente sur deux axes : la valeur de n et la taille des séquences, les taux de vrais positifs et de faux positifs étant utilisés comme critères de qualité.

Classification des promoteurs

Les taux de vrais positifs sont ici comparés pour des valeurs de n allant de 1 à 7 (la taille de l'échantillon d'apprentissage positif ne permet pas d'aller au delà de cette valeur). La figure 5 montre que les ngrams reconnaissent moins de promoteurs.

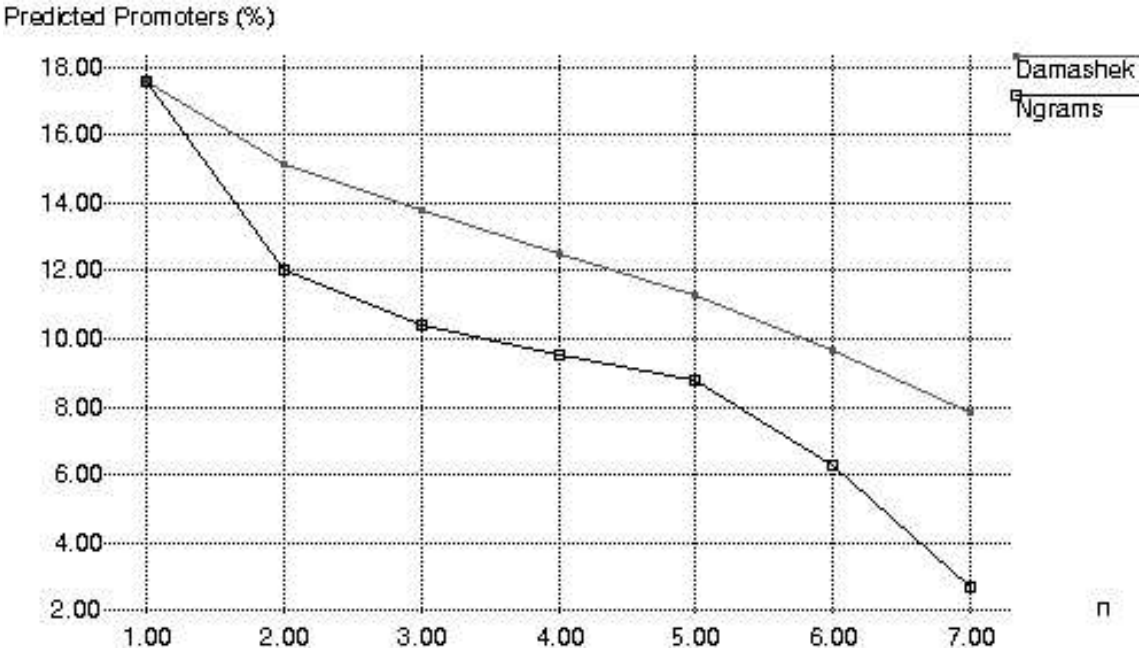

FIG. 6 – *Taux de faux positifs* ($n = 1 \ldots 7$)

Classification des non promoteurs

Ici encore nous considérons des séquences de taille 600. Le taux de faux positifs est plus élevé avec la méthode des fréquences (figure 6).

Ces résultats sont en faveur des ngrams, la perte sur les vrais positifs (-14% pour $n = 6$) étant largement compensée par le gain sur les faux positifs (-35%) (ceci reste vrai pour des tailles de séquence autres que 600). Rappelons que la densité des régions promotrices dans l'ADN justifie que l'on souhaite réduire le taux de faux positifs au maximum.

Influence de la taille des séquences

Nous n'avons pas étudié l'effet de la taille des séquences concernant l'échan-

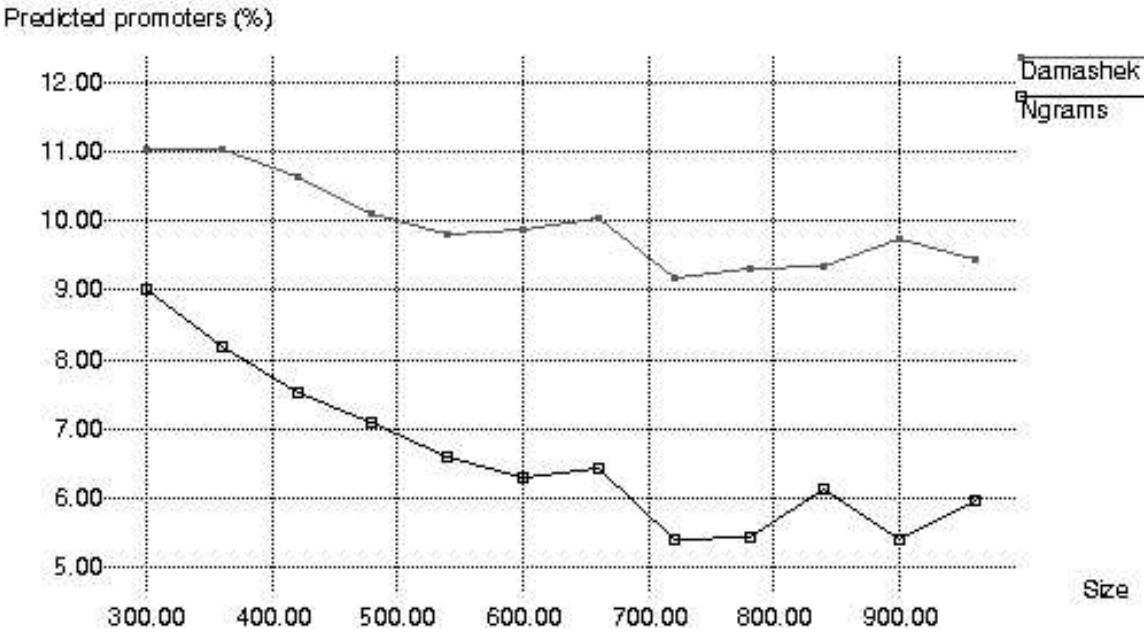

FIG. 7 – *Taux de faux positifs pour différentes tailles de séquence*

tillon de promoteurs, la taille de ceux-ci étant petite et variant beaucoup.

Concernant l'échantillon négatif, le taux de classification correcte pour des séquences de longueur différente fournit une indication sur la fiabilité d'une prédiction en fonction de la taille de la séquence, et permet de fixer la taille de la fenêtre lors de l'exploration d'une grande séquence. Les résultats de la figure 7 concernent des tailles de $300, 360, 420, 480 \ldots 960$.

Comme l'on pouvait s'y attendre les résultats sont meilleurs sur de grandes séquences. Toutefois le taux de faux positifs tend à se stabiliser rapidement, si bien que 700 nucléotides semblent suffire (ceci reste vrai pour d'autres valeurs de n).

En conclusion, cette étude comparative montre que bien que la méthode ngrams

soit basée sur le comptage des n-uplets et que ses résultats dépendent du taux de GC, elle ne se réduit pas à une simple comparaison de la composition de la séquence avec celles des échantillons.

Exemples de détection

La détection de régions promotrices dans une large séquence d'ADN s'effectue en faisant glisser une fenêtre de taille fixée le long de la séquence. Cette fenêtre délimite une sous-séquence dont le score est reporté sur un graphe visualisant les régions potentiellement promotrices.

Les deux séquences étudiées appartiennent à l'ensemble de validation négatif. Sur la première (figure 8) apparaissent clairement trois pics qui correspondent effectivement à trois des cinq débuts de transcription connus de la séquence. Le seul algorithme de détection de promoteurs publié, PROSCAN [Pre95], prédit deux de ces trois promoteurs, plus treize autres non recensés (PROSCAN est basé sur la fréquence de sites de fixation dans un corpus d'apprentissage).

La deuxième, composée à 62,5% de G et C, est reconnue comme composée quasi-exclusivement de promoteurs, alors qu'elle n'en contient qu'un (PROSCAN en trouve 26). Les trois pics visibles sur la figure 9 sont des régions particulièrement riches en GC. On voit ici l'importance pour la reconnaissance des promoteurs de se dégager du biais introduit par la composition des séquences.

3.2 Multigrams

Alors que les ngrams exploitent les dépendances entre un symbole et un contexte de longueur fixe, le modèle multigrams [BPLA94] représente les séquences comme une concaténation de sous-séquences de longueurs variables. La longueur de ces sous-séquences est bornée. On parle alors de n-multigrams si n est la longueur maximale d'une sous-séquence.

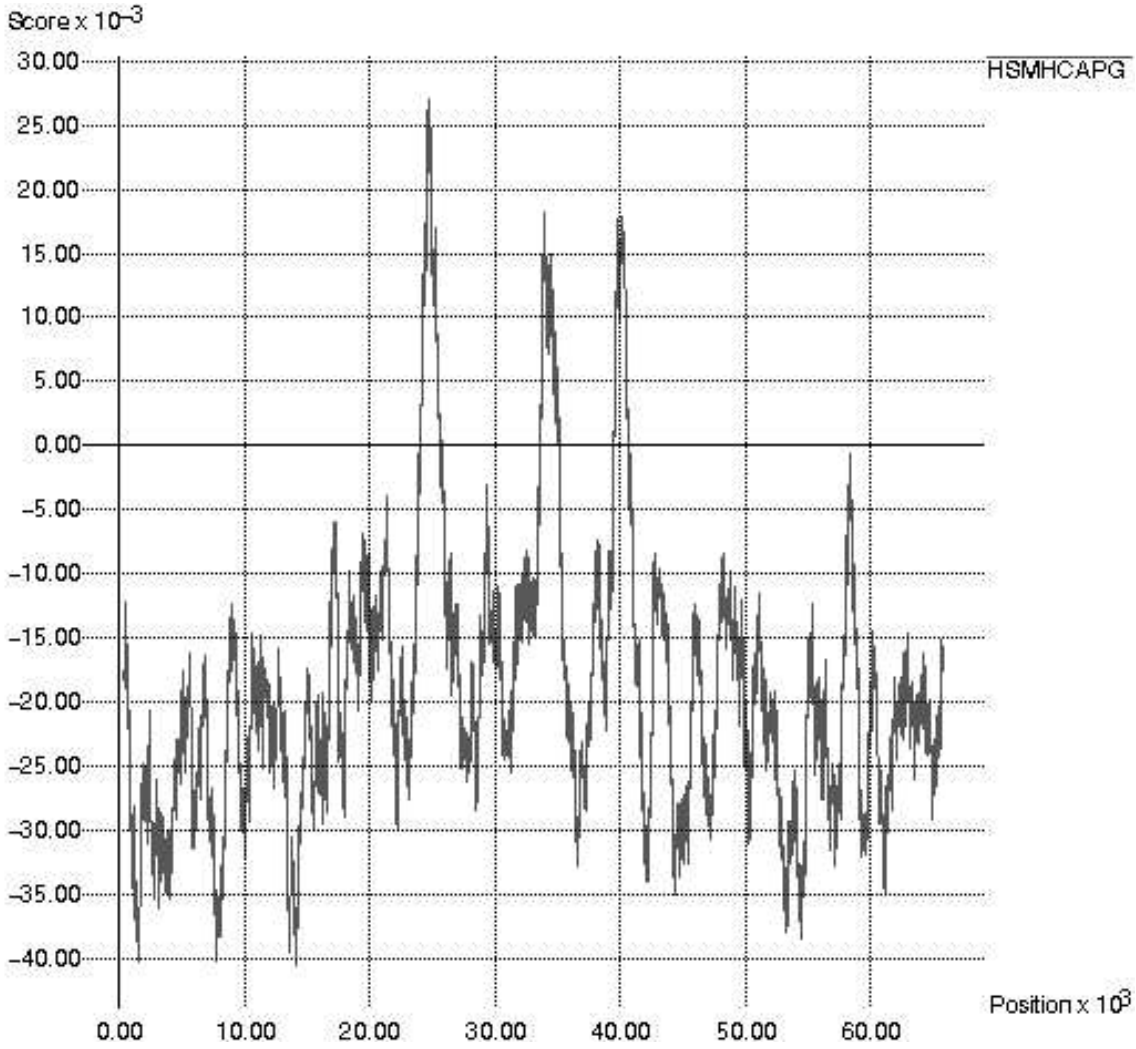

FIG. 8 – *Un exemple de détection réussie*

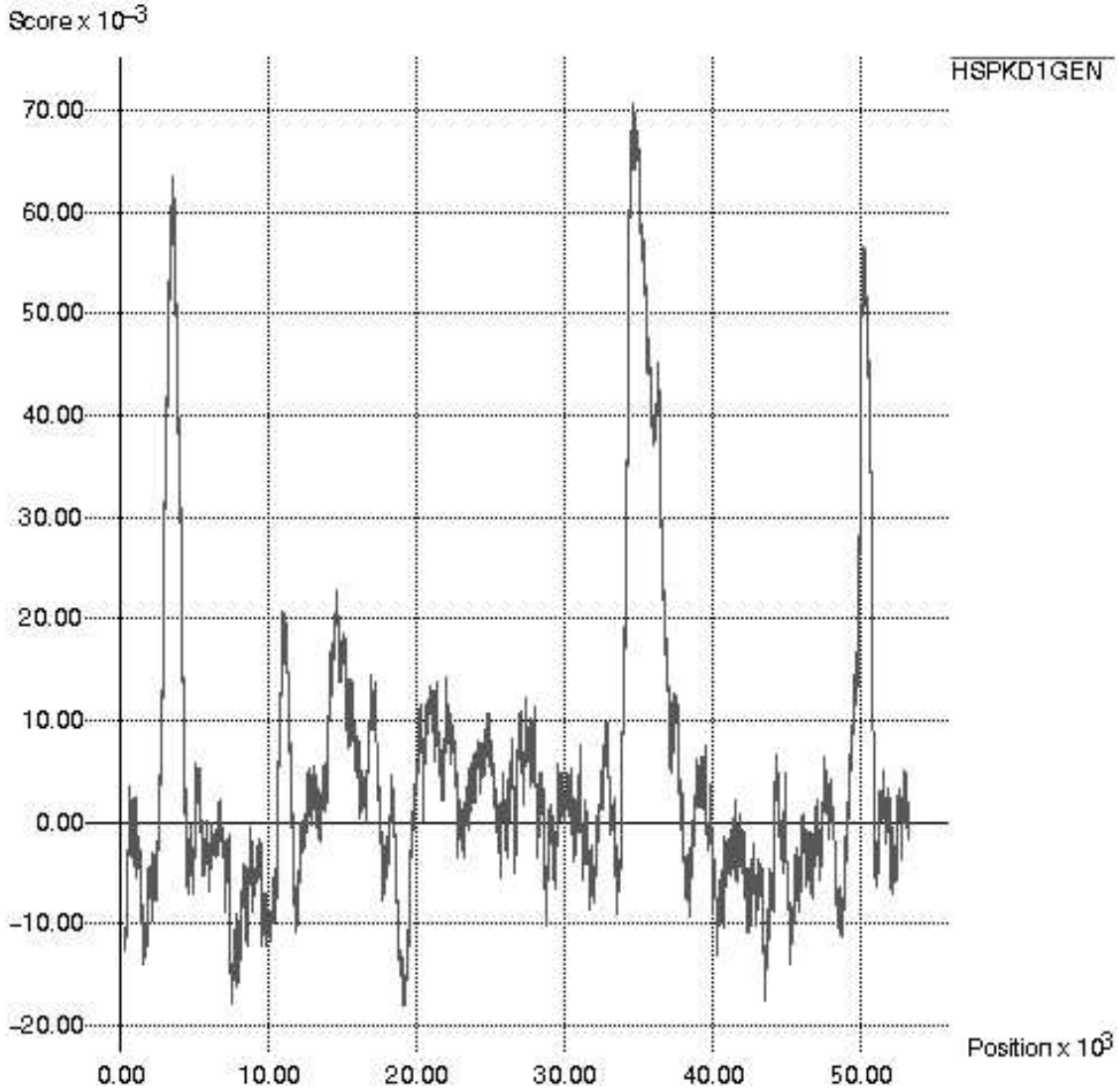

FIG. 9 – *Un exemple d'échec de détection*

3.2.1 Segmentation d'une séquence

Les paramètres d'un modèle n-multigrams sont les probabilités de chacune des sous-séquences. Si on les suppose déterminées, la vraisemblance d'une séquence segmentée est égale au produit des probabilités des sous-séquences composant la segmentation. Plus formellement :

étant donnée une segmentation de $w = w_1 w_2 \dots w_N$ de la forme $w = s_1 s_2 \dots s_q$,

$$\mathcal{L}(w) = \Pi_{i=1}^{q} P(s_i)$$

La segmentation de w s'effectue alors au sens du maximum de vraisemblance, la segmentation optimale pouvant être déterminée par un algorithme de Viterbi.

Ainsi, pour une séquence de 4 symboles et un modèle de 3-multigrams :

$$\begin{aligned}
\mathcal{L}(w) &= \mathcal{L}(w_1 w_2 w_3 w_4) \\
&= \max\{\, P(w_1)P(w_2 w_3 w_4), \\
&\qquad P(w_1 w_2 w_3)P(w_4), \\
&\qquad P(w_1 w_2)P(w_3 w_4), \\
&\qquad P(w_1)P(w_2)P(w_3 w_4), \\
&\qquad P(w_1)P(w_2 w_3)P(w_4), \\
&\qquad P(w_1 w_2)P(w_3)P(w_4), \\
&\qquad P(w_1)P(w_2)P(w_3)P(w_4)\}
\end{aligned}$$

3.2.2 Estimation des paramètres

étant donné un ensemble d'apprentissage constitué de plusieurs séquences, les probabilités des sous-séquences peuvent être estimées en deux temps : une estimation initiale et une estimation itérative. L'estimation initiale des probabilités utilise toutes les sous-séquences présentes dans l'ensemble d'apprentissage, par un simple calcul de fréquences. La ré-estimation s'effectue en segmentant les séquences d'après les probabilités calculées à l'étape précédente. Les fréquences des sous-séquences obtenues fournissent les paramètres d'un nouveau modèle. Ces paramètres permettent de re-segmenter les séquences, le

processus étant itéré jusqu'à convergence.

à l'issue de ce traitement on obtient un dictionnaire de sous-séquences de longueur variable permettant de décrire l'ensemble d'apprentissage.

3.2.3 Application à la reconnaissance de promoteurs

Deux modèles sont calculés, l'un pour les promoteurs et l'autre pour l'ADN. Les vraisemblances de chaque séquence des échantillons de validation sont comparées, ce qui permet d'effectuer une prédiction. Ici encore, les résultats dépendent de la composition des séquences.

Nature des échantillons

En raison de la complexité de la procédure permettant de calculer les paramètres des modèles, nous n'avons pas pu utiliser le même ensemble d'apprentissage négatif que pour les deux méthodes précédentes. Nous avons sélectionné un échantillon de taille réduite ($2,4.10^6$ nucléotides, soit 2684 séquences de 900 nucléotides et un facteur de réduction de l'ordre de 10) disjoint de l'ensemble de validation.

Les autres échantillons sont les mêmes que pour les ngrams, l'échantillon de validation négatif étant constitué de séquences de taille 600.

Résultats

La figure 10 montre les résultats des prédictions pour l'ADN et les promoteurs. Les taux de vrais et faux positifs sont tous deux moindres qu'avec les deux méthodes précédentes. Il est toutefois difficile d'établir une comparaison, l'échantillon d'apprentissage négatif étant différent. Nous avons de plus choisi ici le mode d'apprentissage le plus simple. Il est en effet possible [DB95] de prendre en compte toutes les segmentations possibles lors de l'apprentissage comme du calcul de la vraisemblance d'une séquence.

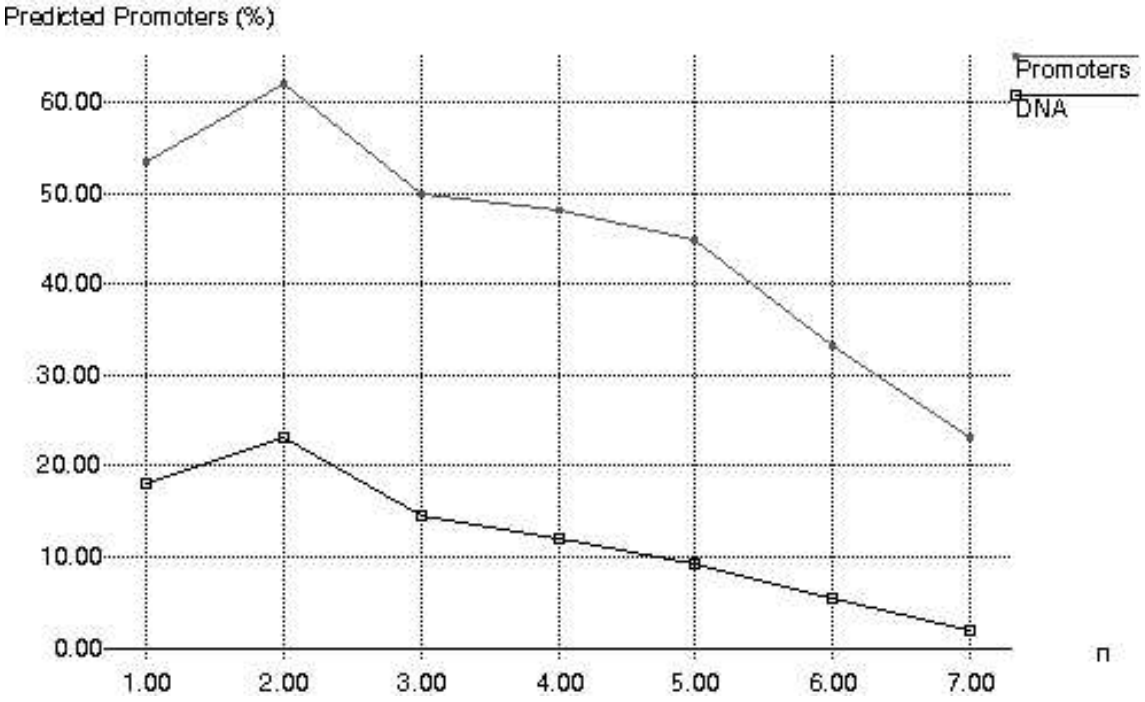

FIG. 10 – *Multigrams : pourcentages de séquences classées promotrices*

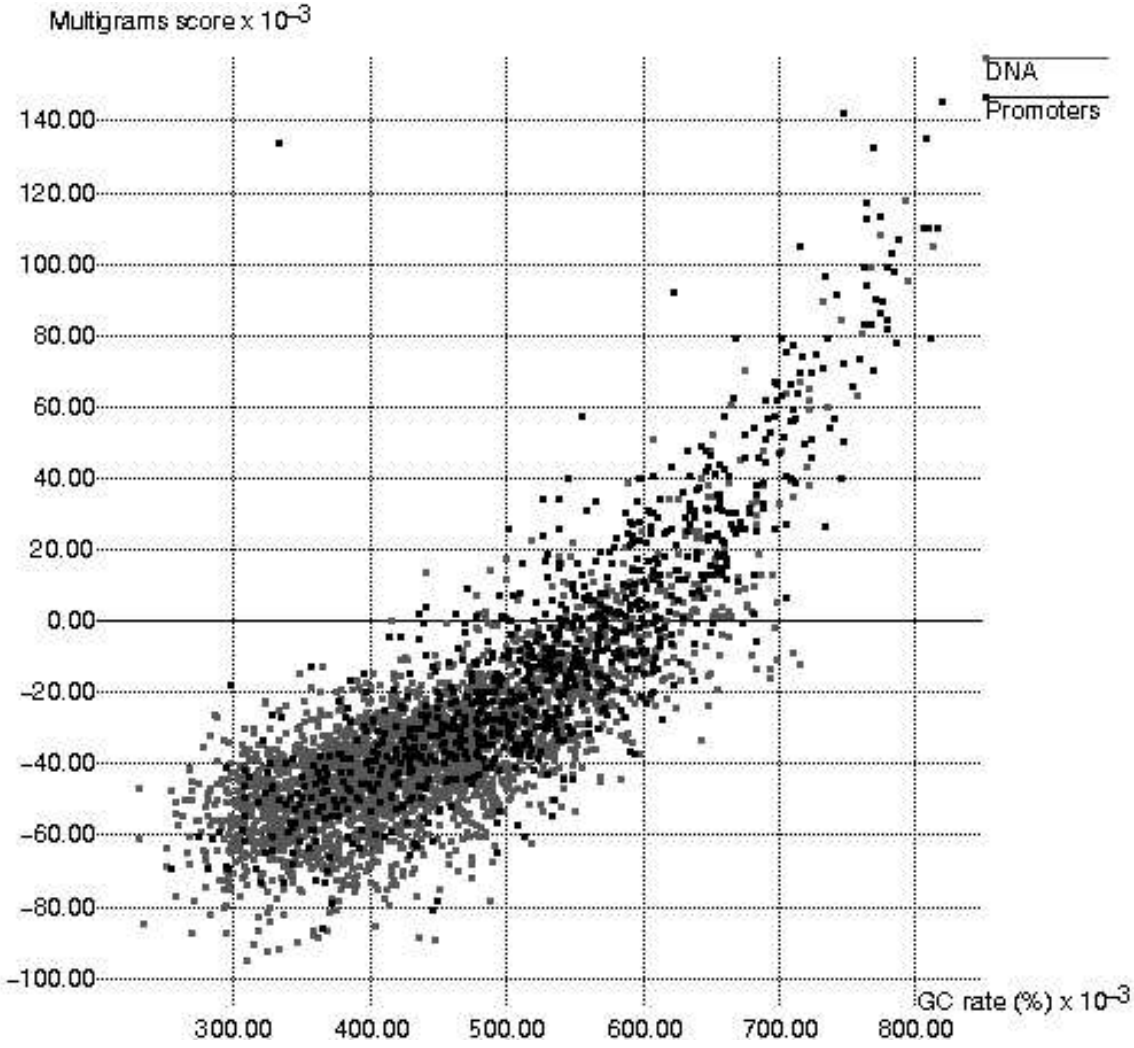

FIG. 11 – *Corrélation Multigrams/Taux de GC*

La modélisation reposant en dernière instance sur un processus de comptage, il n'est pas étonnant que les résultats dépendent encore du taux de GC (figure 11). Nous proposons dans la section suivante une technique tenant compte de la variabilité de la composition des séquences génétiques.

4 Comparaison de z-scores

Nous reprenons ici la méthode de comparaison des fréquences de la section 3.1.3, en rapportant les fréquences observées à la composition des séquences. La notion statistique de z-scores, couramment utilisée en bioinformatique, répond à ce besoin.

4.1 Z-scores

La recherche de motifs caractéristiques d'un ensemble de séquences fonctionnellement équivalentes a fait l'objet de nombreux travaux [Kon94]. L'une des approches consiste à déterminer un ensemble de mots sur ou sous-représentés dans ces séquences. On compare pour cela les fréquences observées aux fréquences théoriques attendues, calculées *a priori* à partir de données statistiques sur les séquences.

La procédure est en général la suivante :

1. Calculer les fréquences réelles $F(i)$ $(i = 1 \ldots 4^N)$ des mots de longueur N.

2. Choisir un modèle à partir duquel les mots considérés pourraient être générés.

3. Calculer les fréquences attendues $F_e(i)$ et leurs variances $V(i)$ relativement à ce modèle.

4. Comparer fréquences observées et attendues, en tenant compte de la variance :

$$z(i) = \frac{F(i) - F_e(i)}{\sqrt{V(i)}}$$

   $z(i)$ est le z-score associé au $i^{\text{ème}}$ mot.

5. Comparer $z(i)$ à une valeur seuil arbitraire z_0. Le mot est considéré comme sur (resp. sous) représenté si $z(i) > z_0$ (resp. $z(i) > z_0$).

4.2 Application à la reconnaissance de promoteurs

La méthode des fréquences de la section 3.1.3 était basée sur la comparaison du vecteur représentant une séquence avec deux vecteurs représentant deux classes. Nous reprenons ce principe, en comparant non plus les fréquences de mots mais leurs z-scores.

4.2.1 Choix d'un modèle de prédiction des fréquences

Nous faisons l'hypothèse que l'observation d'un mot suit une loi binomiale. Ceci est faux, chaque mot déterminant en partie le suivant. Un autre modèle, prenant en compte ces dépendances et plus adapté pour la recherche de motifs [SEGN90], s'est révélé inadéquat, les variances calculées dépendant de la taille des échantillons.

La probabilité d'observation de chaque mot est calculée *via* un modèle ngrams (cf. section 3.1.2) de la séquence : pour $w = w_1 w_2 \dots w_N$,

$$P(w) = P(w_1 \dots w_{n-1}) P(w_n | w_1 \dots w_{n-1}) \dots P(w_N | w_{N-n+1} \dots w_{N-1})$$

La fréquence attendue de w est alors $P(w)$, et sa variance $P(w)(1 - P(w))$.

Différentes valeurs de n ont été testées. Choisir un n trop grand induit une trop grande corrélation entre les paramètres du modèle ngrams et les fréquences observées, si bien que les fréquences théoriques diffèrent peu des fréquences observées. Pratiquement, $n = 1$ et $n = 2$ fournissent les meilleurs résultats.

4.2.2 Résultats

Pour chaque classe de séquences (ADN ou promoteurs), l'ensemble des z-scores de chaque mot peut être représenté par un point dans un espace de dimension 4^N. La distance euclidienne du point représentant une nouvelle séquence, soit S, à chacun des points caractéristiques des deux classes, soient A et P, permet alors de classer la séquence : celle-ci est promotrice ssi $d(S, A) > d(S, P)$.

Notons que la comparaison des z-scores associés aux mots par les deux modèles permet de mettre en évidence la fréquence anormalement élevée dans les promoteurs des deux motifs caractéristiques les plus courants et les mieux conservés : GGGCGG (site Sp1) et TATAAA (TATA-box).

Une étude systématique de l'influence des différents paramètres (n, N, taille des séquences) sur la nature des prédictions reste à mener. Les premières expérimentations sont toutefois probantes, en particulier pour les séquences à majorité de GC. Ainsi pour $n = 2$ et $N = 4$, on a représenté sur la figure 12 le score $d(S, A) - d(S, P)$ associé à une séquence. La méthode discrimine assez bien les séquences comportant plus de 50% de GC (50% de vrais positifs, 10% de faux positifs).

Les deux ensembles d'apprentissage sont ici représentés par deux points. Nous pensons obtenir de meilleurs résultats grâce à une approche du type analyse discriminante, les deux ensembles d'apprentissage étant alors représentés par des nuages de points. Il est toutefois probable que la fréquence anormale de certains motifs n'est pas caractéristique de tous les promoteurs. La majorité des motifs correspondant aux sites de fixation subissent ainsi des altérations d'une séquence à l'autre, ce qui limite notre approche. De plus, l'importance d'un mot dépend souvent de son contexte, au niveau de la séquence elle-même ou des structures d'ordre supérieur de la molécule (conformation dans l'espace).

Quoi qu'il en soit, la section 3.1.4 a montré que la méthode que nous avons adaptée ici n'est pas parmi les plus efficaces. Une voie de recherche est ainsi la conception de modèles plus complexes (ngrams, multigrams...) prenant en compte la composition des séquences.

5 Conclusion

L'utilisation de modèles de langages pour l'analyse de séquences génomiques a été abordée, avec pour but la caractérisation d'un ensemble de séquences fonctionnellement équivalentes, les régions promotrices.

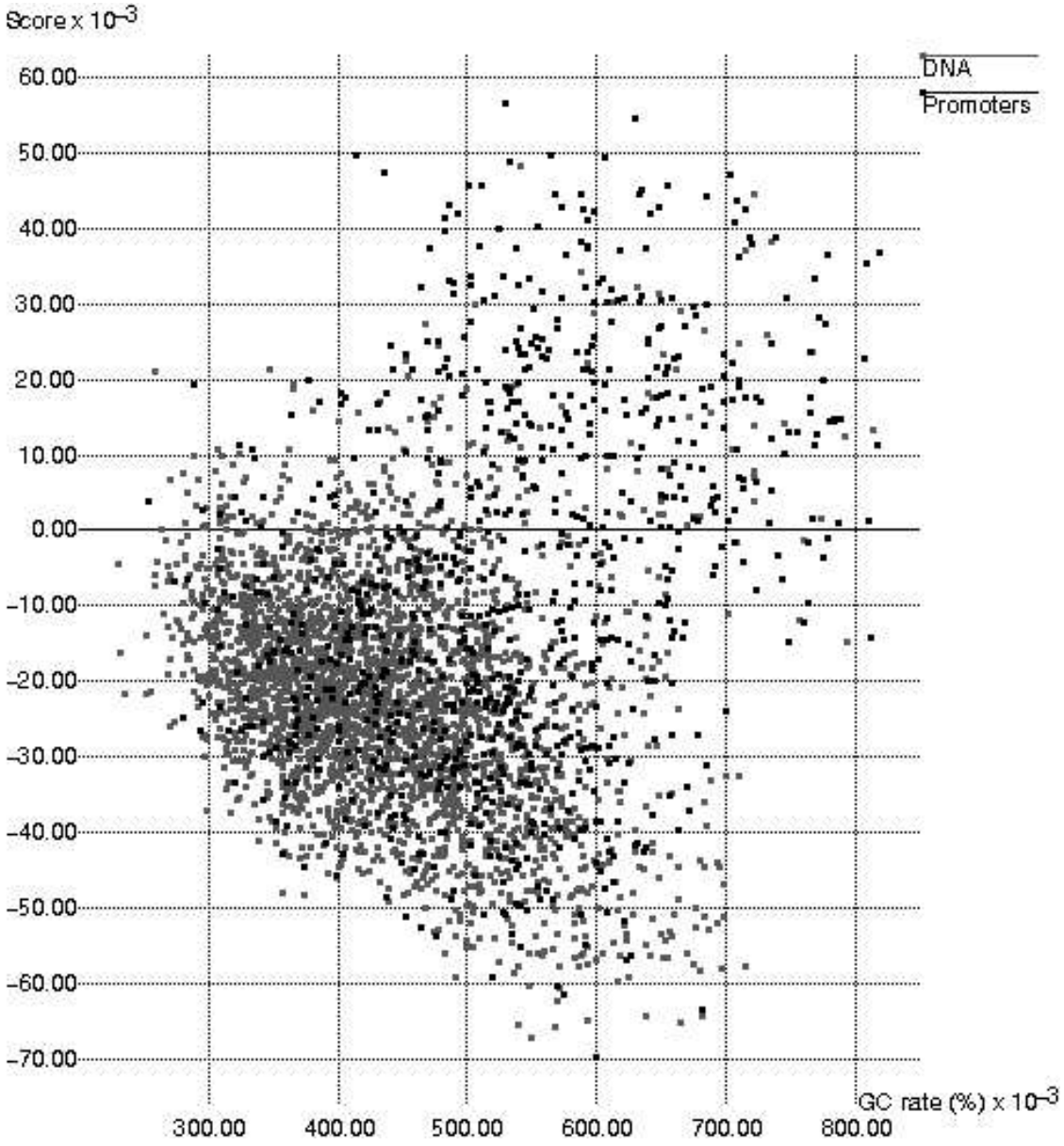

FIG. 12 – *Z-scores et taux de GC*

Un premier travail sur les sites de fixation a débouché sur une nouvelle méthode de détection. L'apport reste cependant mineur, et l'absence de librairies de sites homogènes et exhaustives ainsi que la complexité des mécanismes liant les sites à la régulation nous semblent proscrire une approche basée directement et uniquement sur les connaissances biologiques du phénomène (d'autres connaissances peuvent par ailleurs être exploitées, sur la structure secondaire de l'ADN par exemple [MP96]).

Nous nous sommes ensuite tournés vers la caractérisation de régions promotrices à travers des modèles de langages statistiques. Si ces modèles (ngrams, multigrams) se sont avérés relativement satisfaisants, dans le sens où ils permettent une meilleure détection que les algorithmes dont nous avons connaissance, une grande part de leur efficacité est due à la capture d'informations sur la composition des séquences considérées.

Notre but ultime étant la mise en évidence dans les textes génomiques de structures similaires sinon analogues à celles du langage naturel, nous pensons devoir nous abstraire de cette caractéristique première de la séquence.

Nous avons proposé à cet égard une première méthode fondée sur la comparaison des fréquences des n-uplets dans les deux classes de référence et la séquence considérée, en effectuant une correction en fonction de la composition des séquences. La méthode de comparaison est rudimentaire, mais le système permet, parmi les séquences comptant plus de 50% de GC, de rejeter neuf contre-exemples sur dix tout en gardant un exemple sur deux (ce que ne permettait aucun des systèmes précédents).

Nous comptons dans un premier temps améliorer cette technique, puis poursuivre l'étude des modèles de langages en tentant de les adapter à la disparité de la composition des séquences. Les connaissances ainsi acquises seront à associer aux résultats d'autres approches, par exemple la détection de sites de fixation ou la prédiction de structures d'ordre supérieur.

Remerciements

Ce travail, réalisé au sein de la société GENSET sous la responsabilité scientifique de J. Nicolas de l'équipe Repco, a bénéficié du soutien d'une bourse "postdoctorale industrielle" (convention de collaboration no. 196C0490031310012 entre l'INRIA et la société GENSET).
Adresse de l'auteur :
GENSET, 1, rue Robert et Sonia Delaunay, 75011 PARIS,
Tel : 01 43 56 59 00,
E-mail : Jean-Yves.Giordano@genset.fr, URL : http://www.genset.fr

Références

[AKMSH94] M. Brown A. Krogh, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[BPLA94] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. Modèles de séquences à horizon variable : multigrams. In *Actes des 20èmes Journées d'Étude sur la Parole*, 1994.

[Dam95] M. Damashek. Gauging similarity with n-grams: language-independent categorization of text. *Science*, 267, 1995.

[DB95] S. Deligne and F. Bimbot. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. In *Proceedings of ICASSP*, 1995.

[Dup96] P. Dupont. *Utilisation et apprentissage de modèles de langage pour la reconnaissance de la parole continue.* PhD thesis, ENST, 1996.

[Gio96] J-Y. Giordano. Grammatical inference using tabu search. In *Proceedings of the 3rd International Colloquium on Grammatical Inference*, 1996.

[GSCT94] I. Galiano, E. Sanchis, F. Casacuberta, and I. Torres. Acoustic-phonetic decoding of spanish continuous speech. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(1):155–180, 1994.

[Jel90] F. Jelinek. Self organized language modeling for speech recognition. In A. Warbel and K.K. Lee, editors, *Readings in Speech Recognition*, pages 450–506, San Mateo, CA, 1990. Morgan Kaufmann.

[Kon94] A.K. Konopka. Sequences and codes: fundamentals of biomolecular cryptology. In *Biocomputing*, pages 119–174. Academic Press, 1994.

[Leg95] P. Leguillette. *Analyse informatique de séquences génétiques.* Rapport de stage, 1995.

[Lew95] B. Lewin. *Genes V.* Oxford University Press, 1995.

[Luz94] D. Luzeaux. Process control and machine learning: rule-based incremental control. *IEEE transactions on automatic control*, 39(6):1166–1171, 1994.

[MP96] M. Marilley and P. Pasero. Common dna structural features exhibited by eukaryotic ribosomal gene promoters. *Nucleic Acids Research*, 24(12):2204–2211, 1996.

[OG92] J. Oncina and P. Garcia. Inferring regular languages in polynomial update time. In *Proceedings of the 4th Spanish Symposium on Pattern Recognition and Image Analysis*, 1992.

[Pre95] D.S. Prestridge. Predicting polii promoter sequences using transcription factor binding sites. *Journal of Molecular Biology*, 249:923–932, 1995.

[Sea96] E.R. Sean. Hidden markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.

[SEGN90] E.E. Stuckle, C. Emmrich, U. Grob, and P.J. Nielsen. Statistical analysis of nucleotide sequences. *Nucleic Acids Research*, 18(22):6641–6647, 1990.

[SYSGG95] D. Millinoff S.R. Yant, W. Zhu, J.L. Slightom, M. Goodman, and D.L. Gumucio. High affinity yy1 binding motifs: identification of two core types (acat and ccat) and distribution of potential binding sites within the human β globin cluster. *Nucleic Acids Research*, 23(21):4353–4362, 1995.

[YXMSU94] J.R. Einstein Y. Xu, R.J. Mural, M.B. Shah, and E.C. Uberbacher. An improved system for exon recognition and gene modeling in human dna sequences. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, 1994.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399